Mass spectrometry in high-throughput proteomics: ready for the big time

Tommy Nilsson[1,2], Matthias Mann[3], Ruedi Aebersold[4,5], John R Yates III[6], Amos Bairoch[7,8] & John J M Bergeron[1,2]

Mass spectrometry has evolved and matured to a level where it is able to assess the complexity of the human proteome. We discuss some of the expected challenges ahead and promising strategies for success.

© 2010 Nature America, Inc. All rights reserved.

The age of whole-genome sequencing has made the research field of proteomics possible. Proteomics offers highly complementary information to genomics; as most biological functions are transmitted through proteins, proteomics has yielded new biology and insight into disease. Mass spectrometry (MS) technology allows proteins to be analyzed rapidly, accurately and with high sensitivity at a relatively low cost and, when properly applied, with high reproducibility. With current MS technology, several thousand proteins can be identified and quantified in a single study. As it is estimated that the human genome consists of no more than 20,300 protein-coding genes[1], it seems that quantification of the whole human proteome is within reach.

In contrast to a genome, however, a proteome is seemingly boundless as each protein may be present in different forms, in different amounts and at different times. Because of this great complexity, many believe that quantitative profiling of all proteins expressed in a cell at a particular time may be an unachievable goal. Proteomics has also been viewed with skepticism as it has largely failed to deliver new biomarkers for diseases, an application for which it was highly touted[2]. This was in part the result of early high-profile studies[3] that could not be confirmed and, consequently, have negatively affected the credibility of proteomics. A recent comparative study also demonstrated that 20 out of 27 proteomics laboratories (with varying levels of expertise) had difficulties in identifying proteins in even a simple mixture[4,5].

However, when proteomics technologies are applied carefully and correctly, the technology is highly reproducible. An understanding of these techniques and their limitations is necessary for success and has facilitated the development of improved strategies to overcome problems in proteomics and advance the field[6].

Pipeline quality control

The main problem areas of proteomics can be separated into sample preparation, sample handling, data analysis and data evaluation. Whereas biologists and clinicians often control how samples are generated and stored, mass spectrometrists are typically in charge of all other steps, including evaluation of data obtained. This division of labor can result in serious data quality issues. In our opinion, a lack of accountability and management of the data generated will cause the proteomics field to continue to suffer. Biologists and clinicians should not be left in the dark as to the validity of data obtained. It is neither acceptable nor practical to say that 'someone else should validate what I just produced'. Mass spectrometrists best know the limitations of MS and thus are in the best position to validate the data. If the clinician or cell biologist attempts to determine the validity of the data without knowing the limitations and pitfalls of the technology, then it is not possible to distinguish whether the experiment was successful or not. Through routinely performing checks and balances, similar to those used when purifying a protein by means of classical biochemistry, the mass spectrometrist must ensure high fidelity of the platform. Protein standards to test that trypsin digestion, peptide elution, chromatography and MS are in working order are essential. The assignment of spectra to peptides and proteins should also be carefully checked either manually in small data sets such as those generated from single gel bands or spots, or by using accurate statistical tools in the case of large data sets, before handing over any data to the clinician or the cell biologist, and this handoff must include clear instructions as to the interpretation of the data.

Mass spectrometric technology

There are also inherent challenges in mass spectrometry with regards to sensitivity (dynamic range), reproducibility and comprehensiveness, the three of which are interrelated[6]. However, these limitations are understood and can be overcome, and many tools are available to assess the quality of the data.

Sensitivity (dynamic range). Most commercial mass spectrometers have limits of detection in the low femtomole or attomole range, sensitive enough to detect almost any protein. However, the true sensitivity of MS is modulated by the nature of the sample[7]. Biological samples have a wide range of protein abundances, and mass spectrometers are not well equipped to deal with this wide

[1]The Research Institute of the McGill University Health Centre, McGill University, Montreal, Quebec, Canada. [2]Department of Medicine, McGill University, Montreal, Quebec, Canada. [3]Department of Proteomics and Signal Transduction, Max Planck Institute for Biochemistry, Martinsried, Germany. [4]Institute of Molecular Systems Biology, Swiss Federal Institute of Technology (ETH) Zurich, Zurich, Switzerland. [5]Faculty of Science, University of Zurich, Zurich, Switzerland. [6]Department of Chemical Physiology, The Scripps Research Institute, La Jolla, California, USA. [7]Swiss Institute of Bioinformatics, Geneva, Switzerland. [8]University of Geneva, Geneva, Switzerland.
Correspondence should be addressed to T.N. (tommy.nilsson@mcgill.ca) or J.J.M.B. (john.bergeron@mcgill.ca).

dynamic range. For example, about 35% of all predicted human proteins have yet to be observed reliably by MS. This is partly because peptides do not ionize with equal efficiency, potentially putting some proteins (and their associated post-translational modifications (PTMs)) at a disadvantage in terms of detection. This issue is further compounded for low-molecular-weight proteins and those expressed at low abundance.

However, these technical challenges can be overcome. Sample fractionation and preclearing are both routinely used to improve the dynamic range. Low-abundance proteins, such as the tau protein in spinal fluid[8] or cytokines[9], can be detected in a proteome using targeted proteomics approaches.

Reproducibility. MS is also prone to undersampling. Because the process of acquiring tandem mass spectra is driven by computer-controlled algorithms over the course of a finite scan time determined by ion statistics, undersampling can and does occur when peptide mixtures entering the mass spectrometer are complex. Undersampling means that repeated analyses may not yield exactly the same protein identifications, as a slightly different set of peptides may be sampled in the repeat analysis. However, undersampling can be remedied and reproducibility improved by repeating experiments. Additionally, reducing the complexity of samples with extensive prefractionation can reduce the impact of undersampling and improve reproducibility between analyses. Even so, data-dependent sampling of peptides with current mass spectrometers has a well documented degree of variability[10].

The expression of proteins can be strongly influenced by growth conditions, genetic variability and sample handling; these factors further contribute to variability. It is therefore important to minimize technical as well as biological variability among samples. But technical variability is not easily controlled, as shown in a recent comparative analysis where eight laboratories attempted quantification of a carefully designed test sample containing seven different proteins in plasma using three replicates and standard operating procedures, through a targeted approach called selected reaction monitoring (SRM)[7]. The study showed that the SRM method was reproducible in the sense that the same peptides were detected consistently by the different labs and with good quantitative accuracy; the 25% quantitative variation among labs was better than expected. Even so, this puts a limit to reproducibility and begs the question of how multiple measurements can be handled and how samples can be compared to reveal, for example, differences in expression.

One possible solution to the issue of variability in comparing samples is to use peptides that have a high probability of being observed as internal quality control standards. In a recent comparative study[4] where 27 participating laboratories were asked to identify 20 equimolar human proteins and 22 peptides of 1,250 Da, the participants were also required to submit their raw data into a central repository, permitting all tandem mass spectra to be pooled and analyzed centrally. With this central analysis, the peptides seen by most participants became evident; these peptides were observed by MS irrespective of operating procedures or instrumentation and can empirically be classified as 'high-probability' peptides (that is, if the protein is there, there is a high probability that this peptide will be observed). By comparing individual participants' data with the high-probability peptides of the pooled data, it was relatively easy to reveal discrepancies and to identify the errors made by a particular participant. Differences in ionization, in precursor ion selection algorithms and in upstream fractionation techniques do not yield great variability in terms of identification, at least not for high-abundance proteins. For low-abundance proteins, differences among mass spectrometers may be more decisive.

Another question the field must resolve is not just the reproducibility of MS results, but also potential discrepancies with other technologies. For the SRM reproducibility study mentioned above[7], the absolute concentration values measured for seven proteins in plasma samples differed not just among the laboratories using SRM but also differed from those measured by enzyme-linked immunosorbent assay (ELISA). For example, for C-reactive protein, a variation of 0.31 to 1.8 fmol μl^{-1} was observed by SRM at an amount judged to be 4 fmol μl^{-1} by ELISA. Whether this indicates limitations with commercial ELISA assays or with the MS SRM-based approach is not clear, especially as C-reactive protein is comparatively abundant in plasma.

Comprehensiveness. Although the number of protein coding genes can be predicted in a genome, when, where or indeed whether such genes are transcribed and translated into proteins is not obvious. As noted above, 35% of the 20,300 predicted human proteins have little or no evidence for their expression. Given that there are some 230 different cell types in the human body, each presumably expressing only a subset of the 20,300 predicted proteins, defining when a proteome of a cell type, body fluid or tissue is complete is a very difficult proposition. Analyzing the transcripts using, for example, a genome-wide array does not provide the answer, as the correlation factor between mRNA abundance and protein abundance on a genome-wide scale is very low[11].

However, approaches have been developed to determine when the proteome is 'complete'. By combining fractionation and replicate analyses to collect as much data as possible, it is possible to develop saturation curves: that is to say, as each new experiment is run, the number of new proteins or peptides identified decrease until eventually no new proteins are identified. This approach was used to define large portions of the secreted proteomes of human saliva from the parotid gland and the submandibular plus sublingual glands[12], the proteome of a specific *Drosophila* cell type[13], the proteome of rat lung luminal endothelial cell plasma membranes[14] and, in a landmark study, the complete proteome of the budding yeast *Saccharomyces cerevisiae*[15]. This method is contingent on setting a low false positive rate for peptide identification, as the number of 'new' proteins will continue to grow unless this rate is kept low[16,17]. Low false discovery rates can be aided by using highly accurate mass spectrometers, but without tight control of the false positive rate, the number of false positive proteins could continue to climb until every protein encoded by the genome has been identified[18].

To obtain complete proteomes, the field must address the issue of protein extractability. The use of gel-based isoelectric focusing for protein separation invariably leads to a loss of transmembrane proteins in particular, but as much as 70% of all proteins could be lost through this procedure. Solution-based isoelectric focusing has overcome most of this limitation, but the perception is still that transmembrane proteins may be under-represented. However, a very high degree of comprehensiveness can be obtained for transmembrane proteins. A study that put this to the test was the characterization of the proteome of the synaptic vesicle, where 80% of all membrane proteins in the synaptic vesicles

© 2010 Nature America, Inc. All rights reserved.

were characterized, revealing a density of 130,000 transmembrane domains per μm^2 (ref. 19). It is theoretically impossible to fit many more transmembrane proteins into the lipid bilayer of a synaptic vesicle as there would not be space to surround the membrane-spanning domains with phospholipids.

Quantification

Introducing quantification can avoid many of the pitfalls and caveats described above. By determining the absolute or relative amounts of proteins in the sample, it is possible to place a number on contamination (for example, proteins from blood in a tissue or proteins from other organelles in a purified organellar fraction) and to more easily gauge undersampling and comprehensiveness, as well as sensitivity. Quantification also allows samples to be compared with each other. The dimension of quantity, however, puts a greater demand on the MS-based proteomics pipeline, and several methods have therefore been developed to circumvent this. The advantages and disadvantages of different quantitative approaches have been summarized recently[6].

One label-free approach focuses on ion currents carried by several peptides whose sequences match a specific protein to provide an approximation of abundance[20]. Another label-free method scores the frequency of tandem mass spectra assigned to proteins[21]. In both cases, though, increased accuracy becomes in part a function of time spent on sample analysis on the mass spectrometer, which may extend into the hundreds of hours for even a simple project[22].

Applications to cultured cells and mice of SILAC (stable isotope labeling with amino acids in cell culture) have also proven highly successful[23], and because heavy isotope–labeled rodents are now available, an application to whole-animal experiments is feasible[24]. The merit of more targeted approaches is the focus on highly sensitive and reproducible detection of representative peptides for each protein, such as in SRM. In most cases there are peptide sequences that are unique to each protein of the protein-coding genome. Those unique peptides that also ionize and 'fly' well in the tandem mass spectrometer are referred to as proteotypic peptides[25]. Evaluating the abundance of proteotypic peptides in digests of complex proteins is now feasible by SRM or label-free methods.

In our opinion, by using combinations of quantitative approaches with SRM, it should be feasible to quantitatively map complete proteomes of higher eukaryotes, including humans. This, however, will require highly curated and annotated databases of proteotypic peptides[15,26,27]. The availability of synthetic proteotypic peptide reagents to the community would enormously facilitate the characterization of the entire detectable protein complement of any biological sample[28] although SRM alone is not the answer to obtaining a complete proteome.

Antibody reagents to target proteins of low abundance would also be a very valuable resource for the community. Most studies identify abundant proteins, whereas proteins of low abundance are rarely seen; 50% of all tandem mass spectra generated so far and deposited in public databases only account for 82 human proteins (L. Martens, personal communication). As mentioned above, this can be partially overcome through biochemical pre-fractionation or subcellular fractionation. However, prefractionation is not always sufficient, especially for fluids such as blood where a sixfold difference in dynamic range exists. For low-abundance proteins, an antibody capture method to capture the entire protein and its isoforms has proven both quantitative and reproducible as exemplified by the low-abundance tau protein found in cerebrospinal fluid[29]. The SISCAPA (stable isotope standards and capture by anti-peptide antibodies) method, whereby anti-peptide antibodies are used to capture trypsinized peptides in samples, has been coupled to SRM for absolute quantification[30].

Databases

Protein databases are a necessary part of MS-based proteomics pipelines to identify the mass spectra generated by shotgun approaches. However, data matching and protein naming has been identified as one of the main bottlenecks in MS-based proteomics[4]. The scientific issues with databases have been compounded with political and socio-economical considerations, and the net result is a state of affairs that is far from acceptable. This has become increasingly evident as we put increasing demands on bioinformaticians, including the curators of the databases.

It is important to point out the distinction between 'simple' organisms such as bacteria, archaea and lower eukaryotes as opposed to complex eukaryotes such as *Homo sapiens*, representing the pinnacle of complexity and significance. Most model organisms are endowed with what is termed a model organism database. Model organism databases try to organize and disseminate genetic, molecular and, increasingly, functional information relevant to their species of interest. Paradoxically, because of the intrinsic importance of our species, no single institution has been endowed with the responsibility of centralizing molecular information concerning *Homo sapiens*. The resulting picture for the user is then one of utter confusion because, at present, multiple resources cater to overlapping needs. This could still be satisfactory if it were possible to map 'objects' from one of these resources to the other. Unfortunately, automatic mapping between any two databases rarely results in a success rate above 95%, which is unacceptable. The situation is not helped by the tendency of proteomics users to 'mix and match' the results from different resources such that they obtain the biggest possible set of identifications. The net effect of this is that protein lists full of a hodgepodge of multiple identifiers are generated that often or at least in part are obsolete or irrelevant. One simple way to avoid this should be to only cite protein names. However, this introduces yet another problem because many human proteins do not yet have stable (read single), representative and well defined names.

So is there any hope? Yes, definitely! The US National Center for Biotechnology Information (NCBI), European Bioinformatics Institute (EBI) and Swiss Institute of Bioinformatics (SIB), as well as the HGNC (Human Gene Nomenclature Committee), are working together to come up with a clean set of genes, gene symbols, gene sequences and corresponding protein names and sequences. Already, the Swiss-Prot section of the UniProt knowledge base (UniProtKB/Swiss-Prot) provides a fully annotated and trackable set of 20,300 master protein entries, most of which are linked to a single protein-coding gene. This master set is complemented by more than 14,500 additional splice isoforms, 62,000 annotated sequence variants and more than 80,000 PTMs.

At the level of search engines, the software used to match peptide fragmentation spectra to peptides in the database, there are also problems that must be addressed. Almost no search engines make use of the annotation available in databases such as UniProtKB/Swiss-Prot. Some do take into account splice

© 2010 Nature America, Inc. All rights reserved.

variants, but most search engines have not made any effort to import any information other than the protein sequence. Therefore, they miss out on all annotated and relevant known PTMs and, even worse, fail to take polymorphisms into account. This oversight leads to incorrect or absent identifications.

The HUPO Proteomics Standard Initiative has proposed a common database format, PEFF (PSI Extended FASTA Format) (http://psidev.info/index.php?q=node/363), to store information relevant to the proteomic community. If implemented by commercial and academic search engine developers, this format would address most of the current oversights of search engines. The PEFF format files that will be used by the UniProt consortium will not contain all information stored in UniProtKB/Swiss-Prot, but will provide what is necessary to facilitate the process of identification (for example, known PTMs, polymorphisms and taxonomic information).

Understanding human biology through mass spectrometry

We are convinced that MS-based proteomics, when used with appropriate restraint and properly controlled, is applicable to the study of the human proteome, both on a hypothesis-driven level and on the level of a large-scale effort to comprehensively map the human protein landscape. Proteomics is ready to define the roles of proteins in normal cellular physiology versus disease by accelerating traditional cell biology experiments and enabling new approaches to discovery. Aberrant functions may be dictated by mutations, polymorphisms, isoforms or modification states of proteins; MS-based proteomics is well positioned to measure the types of change that may influence functions and physiology. Proteomics can also define the higher-order organization of cells; that is, where and when proteins localize in the cell as a function of time or disease.

The gathering of MS data with the aim to comprehensively map the human proteome must be done in a way that it does not detract from ongoing activities motivated and driven by scientific curiosity, hypothesis testing and clinically relevant questions. The huge combinatorial space of the human proteome encompasses all body fluids and 230 different cell types. Each cell type and body fluid possesses a unique proteome that depends on which particular genes are transcribed, pre- and post-translational regulation, post-translational processing, protein sorting, intracellular localization, endocytosis, exocytosis, and protein and membrane degradation, all in combination with isoforms, splice variants, amino acid–relevant single-nucleotide polymorphisms and more than 200 different PTMs. Hence, the construction of a human proteome resource must be a long-term process.

In our opinion, it is possible to proceed in much the same manner with which we today conduct basic or clinical research. The main difference would be to pay explicit attention to using published data in a more efficient manner. Here, a gene-centric human proteome may serve as a useful scaffold, as in the building of a skyscraper. Its foundation, the human genome, has already been laid[31–33], as has most of the building itself, through neXtProt, a comprehensive knowledge resource on human proteins, hosting some 20,300 rooms (protein-encoding genes) all ready to be furnished or already partially furnished. The gene/protein–centric nature of neXtProt will ensure that each room is populated such that all associated variants and PTMs, spatial and temporal coordinates are collated, along with their functions and relevance to disease.

Another challenge is then to ensure that the data accumulated are made available in an intuitive and useful manner. This would necessitate the creation of a repository for raw MS data but would ensure that data from MS-based studies have long-lasting value. The ProteomeXchange consortium members, which include Tranche[34], PeptideAtlas[35], the NCBI peptidome[36] and PRIDE[37], already go a long way as potential repositories for a human proteome. Providing these are sufficiently annotated, such a continuously growing data set could then be 'diced' through centralized analysis such that all that is known for a particular protein, process, structure, cell type, organ or condition, and disease at that point in time may be revealed instantly through the push of a button. This, however, does not preclude targeted large-scale efforts[38–40], and indeed, neXtProt already connects to genome-wide projects such as the Human Protein Atlas project[41], which aims to map the expression and localization of proteins in human tissue using immunohistochemistry and immunofluorescence. Also, data from complementary efforts, such as genome-wide RNA interference screens[42] and GFP fusion protein localizations[43], can be incorporated into neXtProt. Chemical genomics information can be included; phenotypic perturbations can be used to explore the molecular basis for alterations in phenotype using MS-based proteomic methods. Further elements of the cell's higher organization need also to be included; proteins associate with other proteins to form operational enzyme or signaling complexes as well as super-structures such as chromatin, nuclear pores or kinetochores. These are not static structures, and thus how protein interactions are regulated as a function of physiology, both normal and abnormal, must be defined to better understand the underlying cell biology.

Many of the questions of cell biology can be addressed through traditional approaches, but perhaps can be answered more quickly and definitively with MS-based methods[44]. Other types of questions have been enabled by the creation of large-scale proteomic methods for protein expression or modification profiling. These methods are now poised to address how expression or modifications change as a function of disease. Sequencing the human genome was perhaps the easy part, and now making sense of the constantly moving and changing picture of the proteome will require a lot of time, effort and creativity.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Clamp, M. *et al. Proc. Natl. Acad. Sci. USA* **104**, 19428–19433 (2007).
2. Anonymous. *Nature* **452**, 913–914 (2008).
3. Petricoin, E.F. *et al. Lancet* **359**, 572–577 (2002).
4. Bell, A.W. *et al. Nat. Methods* **6**, 423–430 (2009).
5. Mann, M. *Nat. Methods* **6**, 717–719 (2009).
6. Domon, B. & Aebersold, R. *Nat. Biotechnol.* **28**, 710–721 (2010).
7. Addona, T.A. *et al. Nat. Biotechnol.* **27**, 633–641 (2009).
8. Westman-Brinkmalm, A. *et al. Front. Biosci.* **14**, 1793–1806 (2009).
9. Nettikadan, S. *et al. Mol. Cell. Proteomics* **5**, 895–901 (2006).
10. Tabb, D.L. *et al. J. Proteome Res.* **9**, 761–776 (2010).
11. Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S. & Weissman, J.S. *Science* **324**, 218–223 (2009).
12. Denny, P. *et al. J. Proteome Res.* **7**, 1994–2006 (2008).
13. Brunner, E. *et al. Nat. Biotechnol.* **25**, 576–583 (2007).
14. Durr, E. *et al. Nat. Biotechnol.* **22**, 985–992 (2004).
15. de Godoy, L.M. *et al. Nature* **455**, 1251–1254 (2008).
16. Tabb, D.L. *J. Proteome Res.* **7**, 45–46 (2008).
17. Brusniak, M.Y. *et al. BMC Bioinformatics* **9**, 542 (2008).
18. Reiter, L. *et al. Mol. Cell. Proteomics* **8**, 2405–2417 (2009).

© 2010 Nature America, Inc. All rights reserved.

19. Takamori, S. *et al. Cell* **127**, 831–846 (2006).
20. Silva, J.C., Gorenstein, M.V., Li, G.-Z., Vissers, J.P.C. & Geromanos, S.J. *Mol. Cell. Proteomics* **5**, 144–156 (2006).
21. Blondeau, F. *et al. Proc. Natl. Acad. Sci. USA* **101**, 3833–3838 (2004).
22. Gilchrist, A. *et al. Cell* **127**, 1265–1281 (2006).
23. Krüger, M. *et al. Cell* **134**, 353–364 (2008).
24. McClatchy, D.B., Dong, M.-Q., Wu, C.C., Venable, J.D. & Yates, J.R. III *J. Proteome Res.* **6**, 2005–2010 (2007).
25. Kuster, B., Schirle, M., Mallick, P. & Aebersold, R. *Nat. Rev. Mol. Cell Biol.* **6**, 577–583 (2005).
26. Picotti, P. *et al. Nat. Methods* **5**, 913–914 (2008).
27. Malmstrom, J. *et al. Nature* **460**, 762–765 (2009).
28. Picotti, P. *et al. Nat. Methods* **7**, 43–46 (2010).
29. Portelius, E. *et al. J. Proteome Res.* **7**, 2114–2120 (2008).
30. Whiteaker, J.R., Zhao, L., Anderson, L. & Paulovich, A.P. *Mol. Cell. Proteomics* **9**, 184–196 (2010).
31. Collins, F.S., Green, E.D., Guttmacher, A.E. & Guyer, M.S. *Nature* **422**, 835–847 (2003).
32. International Human Genome Sequencing Consortium. *Nature* **431**, 931–945 (2004).
33. Gregory, S.G. *et al. Nature* **441**, 315–321 (2006).
34. Hill, J.A., Smith, B.E., Papoulias, P.G. & Andrews, P.C. *J. Proteome Res.* **9**, 2809–2811 (2010).
35. Desiere, F. *et al. Nucleic Acids Res.* **34**, D655–D658 (2006).
36. Ji, L. *et al. Nucleic Acids Res.* **38**, D731–D735 (2010).
37. Jones, P. *et al. Nucleic Acids Res.* **34**, D659–D663 (2006).
38. Faca, V.M. *et al. PLoS Med.* **5**, e123 (2008).
39. Sun, A. *et al. J. Proteome Res.* **9**, 50–58 (2010).
40. Yoshida, Y., Miyamoto, M., Bo, X., Yaoita, E. & Yamamoto, T. in *Proteomics in Nephrology — Towards Clinical Applications* (ed. Thongboonkerd, V.) 186–197 (Karger, 2008).
41. Pontén, F. *et al. Mol. Syst. Biol.* **5**, 337 (2009).
42. Neumann, B. *et al. Nat. Methods* **3**, 385–390 (2006).
43. Starkuviene, V. *et al. Methods Mol. Biol.* **457**, 193–201 (2008).
44. Bergeron, J.J., Au, C.E., Desjardins, McPherson, P.S. & Nilsson, T. *Trends Cell Biol.* **20**, 337–345 (2010)

© 2010 Nature America, Inc. All rights reserved.